Aalto University
School of Business

Congestions and big data

A review on the predictive solutions

Bachelor's Thesis
Juho Ylimartimo
23.5.2017
Business Technology

Approved in the Department of Information and Service Economy xx.xx.20xx
and awarded the grade

TABLE OF CONTENTS

1. Introduction to the subject

1.1 Forewords

At first, cars offered travelling at velocities yet unseen. However, they soon introduced people to crashes, vehicular emission smogs etc. The example shows that issues with new technologies are often unexpected. A notable example of this are traffic congestions (or gridlocks), slowdowns in the service rate of a road traffic network when the service demand goes up (fig. 1).

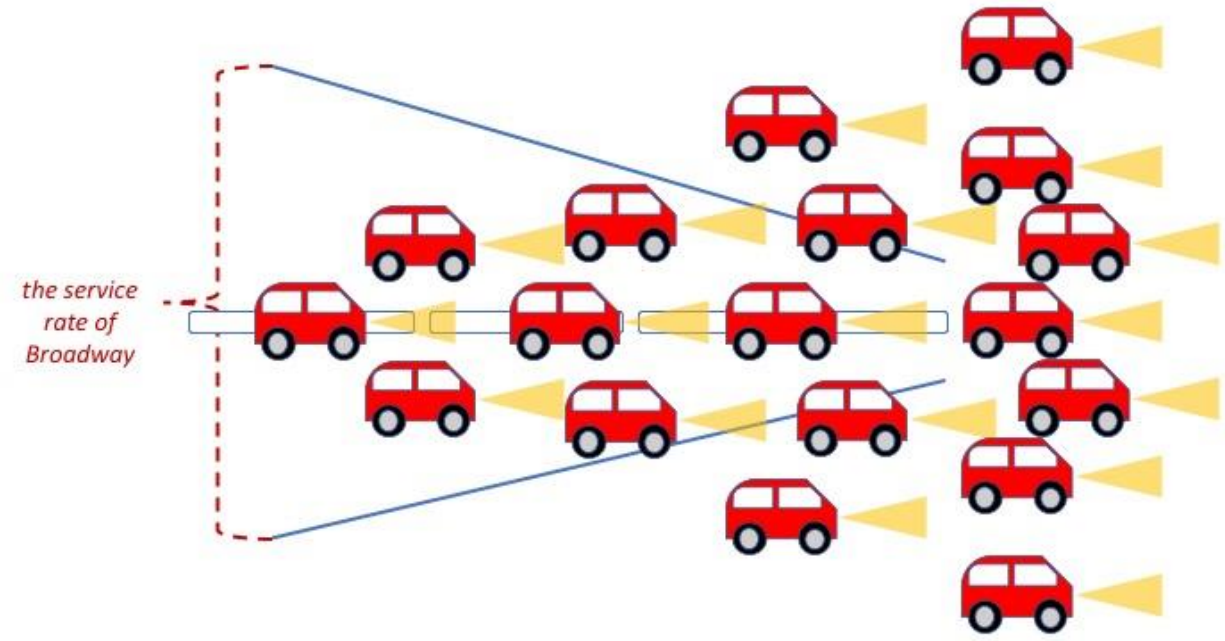

Figure 1. Traffic congestions are slowdowns in the service rate of a road traffic network

Road traffic network that is not sufficient for the demanded service level congests constantly. Thus, it needs to be reengineered and possibly entirely demolished. However, construction plans are specific; i.e. there are no ready solutions that are just installed on their place. While there are no low-cost solutions, traffic flow (the counterforce of traffic congestion) can be enhanced by applying “big data”.

While the causes of traffic congestions may be in part due to physical structures, in this literature review those will be ignored. However, traffic congestions can be remedied (to some extent) with the ambiguous big data, or in short large pools of unstructured data. Unarguably the matter has business worth, or i.e. great savings in

costs that congestions cause.

Therefore, the research questions of this literature review are:

1) How the big data relevant to traffic congestions is collected.
2) How it is processed with appropriate algorithms for information.
3) And how the information is communicated to drivers at the present (the available commercial solutions).

The course of this literature review is as follows: first (1.) in this chapter some recurring concepts are explained. Following up (2.) is the literature review on subject "predicting traffic congestions by big data". And finally (3.), solutions are presented.

1.2. About congestions, and big data

To conclude why big data is the next step in forecasting traffic congestions, a previous innovation is reviewed: toll roads. They do not forecast traffic congestion, but rather they aim to buffer them through market mechanism. As they pose the contrary to the methodologies presented here, they emphasize the issue well.

Toll road's users pay the bill of transporting people safely and fluidly across distances. Hence, the road is a commodity: if not satisfying, then don't buy. Ideally, the only ones using the road are those in need, and hence there are no traffic congestions. In other words, traffic congestions are mismatches between supply and service demand (Abdel-Aty, Shi, 381, 2014).

Market theory says that traffic congestions can be overcome by pricing. Yet, in some of the largest modern metropolitan areas there are 10s of millions of people in traffic. As the theory dictates, price of transportation should come close to a tremendous sum and much lost worth. Hence, direct action against congestions is needed.

Taking these actions requires understanding congestions. I.e. data relevant to the issue is needed to build models on traffic congestions. The question is, what data is there and which is relevant? One did not need to think about this too much before "big data" started to pool up.

Big data is a term coined up in 1990s. It refers to immense pools of

unstructured data. These can be put in context through advanced algorithms. Big data pools up in several steps starting with sensors. These sensors are not just about someone e.g. submitting to an opinion poll in internet, but all kinds of sensors to various metrics such as thermometers.

Should these sensors have access to internet (internet of things, or IoT) they can make data transmissions into a database hosted in cloud. The term refers to a database that is brought to customer through internet, with the physical database maintained in a place across distance.

Clouds are non-siloed which means that previously separate databases converged to offer unprecedented volumes of data. This is a key characteristic of big data. The relationship between these techs is visualized in the chart below (fig. 2):

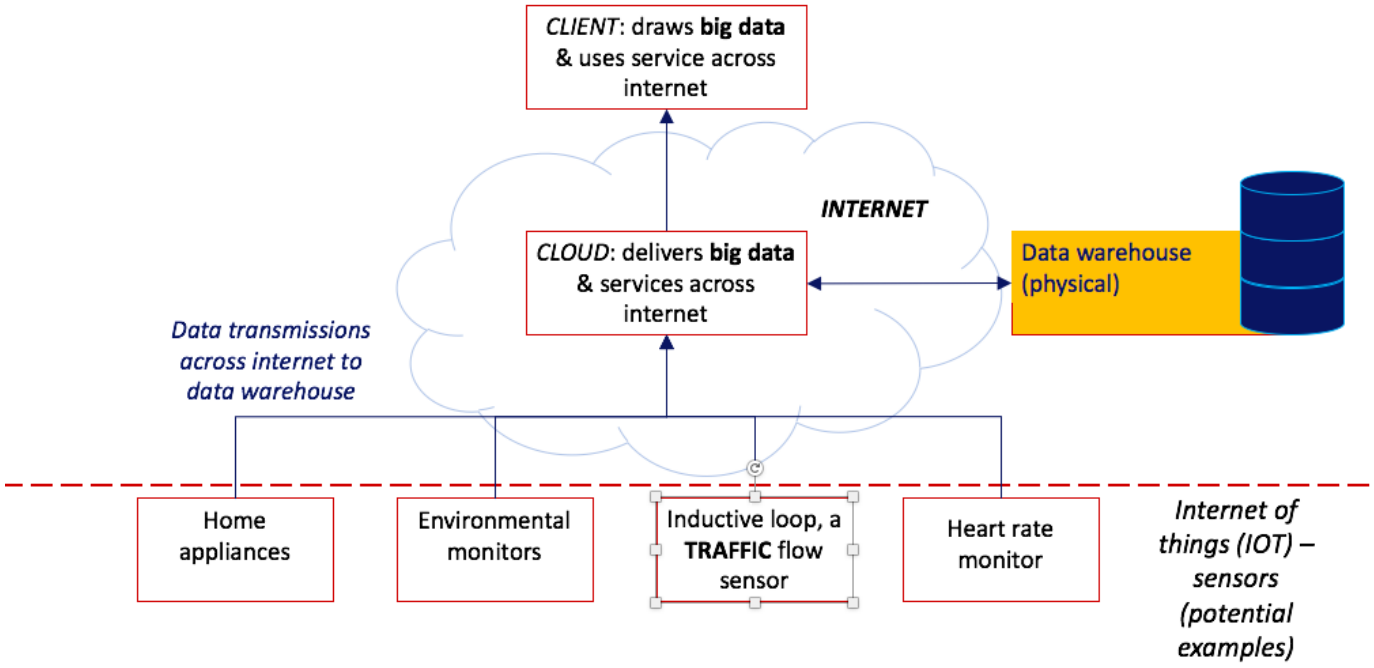

Figure 2. Big data is a technology that is enabled by inventions such as cloud and IoT. In this chart arrows show the direction of data transmissions. Cloud receives these from IoT sensors and submits them to a data warehouse. Then, it distributes data from there to client's purposes. The purpose here is to predict congestions. To add, multiple sensors are used and hence "big data".

As demonstrated, big data is enabled by various inventions. Therefore, in this thesis big data is referred to as algorithms and analytics. The next subsection (1.3) discusses big data's commercial worth in predicting and pre-empting traffic congestions. There are several studies that show this worth to be considerable.

1.3 Big data fueled models on traffic congestions

The idea to use data modelling against congestions is not new. Big data perspective started to recur in literature at turn of 2010s, but "future applications" (such as dynamic route guidance) have been mentioned in studies in 2005 (Li, 41-42, 2005). Clearly, smooth transportation is a commercial concern. Indeed, a leading traffic consulting agency INRIX has predicted for total congestion costs of US and Europe combined to rise to 300 bn$ by 2030 (INRIX, 2014).

There are several factors behind these congestions costs. When income level rises, so does the number of cars per household. This is a well-known pattern in developing countries and in addition, these countries urbanize rapidly. Harbin of China gets 400 additional cars a day as combined result of the listed factors (Cui, Liu et al., 4, 2015).

As a result, there are "mismatches" between roads and service demand, and researchers are looking for data-based methods to find them (Cui, Liu et al., 15, 2015). As an article states: urbanization is a process in which big cities lead to big problems that require solutions of big data level (Zheng et al., 38:2, 2014). Traffic congestions' characteristics fulfil this process description.

To emphasize the issue of congestions on theory level, they are presented below as a non-linear phenomenon. The graph on the next page captures an adverse development in the service rate of an imaginary intersection. On the horizontal axis is service demand (cars per minute), and vertical axis points to the respective service rate.

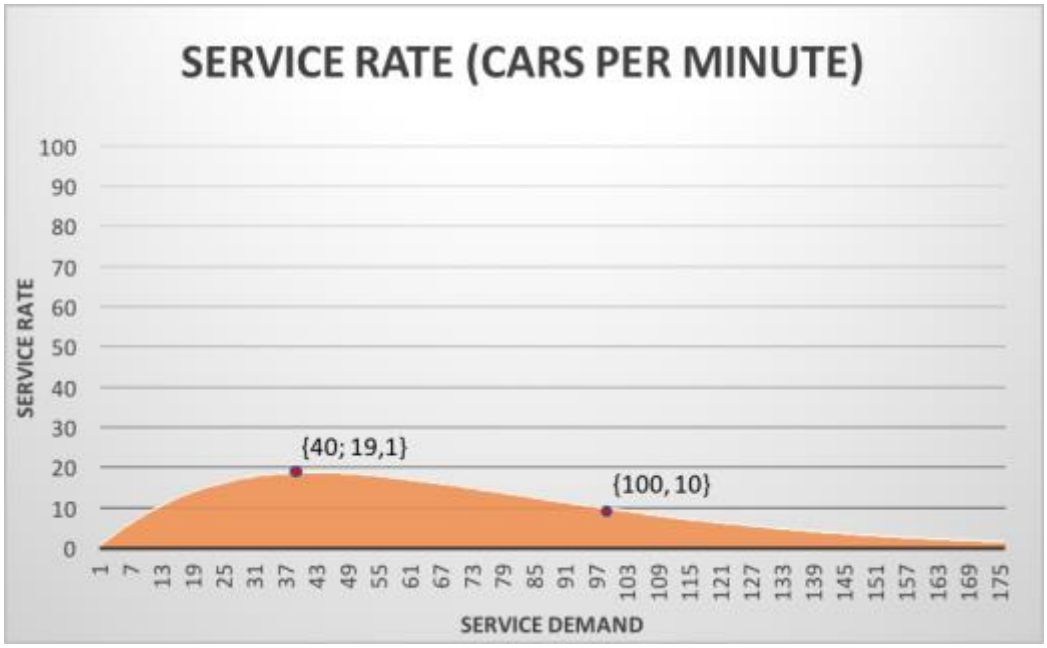

Figure 3. Demonstration of the non-linear relationship between service demand (x-axis) and service rate (y-axis).

The intersection in figure 3 serves 19 cars per minute at service demand of 40. Driver who arrives to the intersection spends roughly 2 minutes in. People may count some allowances in when they need to be on time. However, when demand goes up to 100, and service rate drops to approx. 10 per minute, then each driver waits must wait for 10 minutes.

As demonstrated, traffic congestions are with increasing evidence nonlinear and dynamic (Li, 41, 2005). As in the figure 3 above; 2,5-folds service demand may result in 5-folds wait. In business sense this wasted time is lost monetary worth: missed events, wasted fuel etc. Hence, the solutions to predict congestions may multiply in commercial value as urbanization progresses.

These future solutions quantify and predict traffic congestions and thus preempt their costs. Problem is that traffic congestions are dynamic. E.g. multiple collisions cannot be forecasted, and they can hamper traffic flow for hours. These patterns (such as multiple collisions) must be studied.

Consequently, to study patterns big data is needed. In the volume sense, there is already a plenty of big data, such as all that gps' generate daily. However, according to the literature "big data" –expression is not justified by volume only. Instead, the analysis should have multiple perspectives or different sources of data (Li, Su et al., 292, 2015).

There are many pieces of empiric evidence to support big data's definition in analytics as multiperspective data. As the study in Harbin found out, using taxi GPS data to identify congestions discriminates against certain geographic areas, e.g. shopping districts where taxis are not allowed (Cui, Liu et al., 16, 2015).

Perhaps more significant is that taxi GPS data could also discriminate against certain demographics: for instance, low-income people living in densely populated areas. Low income equals few taxis, and consequently few data. This highlights that traffic congestions cannot be unveiled by naive methods (just one data source).

To repeat, coverage in Harbin's taxi method would not increase by "increasing the number of taxis". Instead, big data means thinking out of the box and unorthodox combinations of data sets. This is demonstrated in the next paragraph.

In China, poor infrastructure prevents traffic flow monitoring in young immigrant areas (Cui, Liu et al., 16, 2015). However, to map out daily traffic flow patterns in those areas, sales of every-day goods could be monitored. This could be a good indicator of business activity, taken that people shop mainly after work. In turn, business activity is at its highest during evening rush hours as people commute.

To add, combining sales data of goods (as demonstrated above) with traffic flow data from is an unorthodox combination. Therefore, big data means data with multiple perspectives to the issue.

1.4 Recurring theories in the literature

For better reading experience, some recurring theories in the literature need to be explained. Predicting traffic flow is mostly about finding causalities. Therefore, the literature uses many terms that come from statistical theory:

1) Regression analysis estimates how one variable affects the other:
    a. *for instance: when service demand on road rises, service rate slows down and congestion gets worse =>*
    b. *the results of regression model are tested by collecting data. The model is "robust" if it does not generate much error*
    c. *Granger causality (a regression model) has been used in the literature to form predictive traffic flow trends.*

2) Normal distribution (or “Gaussian curve”) describes phenomena in which average dominates the whole, e.g.:
    a. *average congestion (in costs) +/- 1 σ (standard deviation) account for roughly 70 % of all costs =>*
    b. *thus, pre-empting average congestions pre-empts most of the costs*

Whether congestion costs are normal distributed or not is essential for the used model’s performance. If prediction model uses Gaussian curve, the impact of average congestions may be overestimated. This may be the case if the true distribution is:

3) Skewed distribution:
    a. *congestions (in costs) are distributed around the average asymmetrically*
    b. *tails may be “fat”: the remaining area after removing data 2 σ from the average > 30 % of the whole =>*
    c. *pre-empting average congestions may not pre-empt costs effectively enough*

This literature review is not about whether congestions follow Gaussian curve. Instead, it should present big data sources such as social media that complement the old forecasting methods. The old methods rely mainly on normal distribution. Therefore, they are reviewed first (2.2.1) and social media data comes after (2.2.2).

The course of the literature review coming up is as follows. First in subsection 2.1 will be introduced the first-hand inventions that are needed to have big data analytics in transportation. Then (2.2) a look is taken at the algorithm (how data sets are used to unravel congestions). For last (2.3) is looked at the present commercial solutions that create value to drivers.

2.1 Intelligent transportation systems, the physical toolbox

Before going to traffic congestions, a recurring yet ambiguous matter in the literature should be defined: the intelligent transportation systems (ITS). ITS are systems in which information techs enhance transportation. I.e., ITS are all the machines that make transportation more efficient and informative to transporters. For instance, traveller information systems (ITS) such as displays in metros give time of departure (Lv et al., 1, 2015).

However, not only are ITS displays and navigator, they are also an important source of big data. In this thesis' case, ITS inductive loop[1] –machines provide traffic flow data to fuel congestion forecasts (Lv et al., 1, 2015). For clarification, the ITS studied in this thesis are mostly like them: traffic flow sensors. Therefore, each time "ITS" are mentioned they refer to infrastructure embedded traffic flow sensors, and vice versa.

Inductive loops are machines that sense traffic flow physically. They fuel congestion forecasts with their big data. As traffic flow sensors (such as inductive loops) are a somewhat old technology to collect data for congestion predictions, their findings are handled first (2.2.1). Non-ITS sources (eg. social media) would be to purposes other than traffic. While so, the presented big data algorithm can also mine their data for traffic related information (2.2.2).

With the algorithm mentioned some clarifying is needed. As described, big data are not the physical databases. Instead, it means enhancing the value of multiple data sources. The big data algorithm are the rules to decide if a source is relevant to the congestions' issue. In the next subsection (2.2) this algorithm is studied.

2.2 The algorithm, or how to teach machine to read and think

ITS are machines that enable big data analytics in traffic. Yet, in their analytic

[1] Inductive loop is a machine embedded in road that registers passing cars using magnetism

capabilities, they are different from humans. ITS are machines, and they are capable of reasoning to the extent they can learn from data. Thus, to learn from data they rely on their programmed rules, or algorithms. The algorithm of this thesis extracts congestion prediction from data.

Take for instance the data from traffic flow sensors: brief rows of data on each passing vehicle such as "@19:09:46.04, weight 1,6 tons". As a road may serve tens of thousands of drivers daily, its sensors receive too many inputs for a human to process. Therefore, "real-time implemented" model must be machine run or i.e. big data techniques applying (Ozbayoglu et al., 1807, 2016).

To emphasize the previous, most of the data defined as big data is bulk. Its primary purpose is not to cater to e.g. congestion forecasting, and it's not necessarily collected by "a congestion forecasting agency". A study used hundred million rows of traffic flow data in an area for one year. This was all non-usable unless put through an algorithm (Ozbayoglu et al., 1807, 2016).

Therefore, ITS are taught to understand data (text) outputs generated by traffic flow sensors. The first input (time, car count) can be outputted to algorithm's next level. Then, this multiple-layer process goes on until we arrive to a traffic congestion prediction, a sort of deep learning algorithm (Lv el al., 868, 2015).

To repeat, classifying between e.g. time, car count etc. in text big data is impossible to do efficiently for human workforce. Thus, it all comes down to how machine is programmed to do multiple-level tasks: from classifying text data inputs to making high-level conclusions how congestions work. Again, this programmed set of rules is an algorithm.

The algorithm of this thesis should probably have the deep learning architecture. Several already exist to fulfill the need for analytics in traffic. However, there is no "one" universal algorithm, but case-specific ones (Lv et al., 866, 2015). Therefore, in this thesis is sought a middle ground with an ideal algorithm that fuels an ideal ITS.

This ITS provides real-time information on the next traffic congestion. As big data sets are often enormous, making these analyses would require tons of human workforce. Thus, this ITS relies on the forethoughtfulness of its algorithm, or i.e. how well it can reason regardless of data's imperfections.

In rough, studies find two different perspectives to viable algorithm: the other is focused on predictions by big data from traffic flow sensors (handled in 2.2.1), and the

other on perhaps more imaginative sensors such as social media (2.2.2). To repeat, a connection between these two is sought.

2.2.1 Congestion by time, the intraday-trend as the basis of congestion prediction

Mining data for congestion forecasts from traffic flow sensors is rather commonplace already. However, big data perspective means recognizing a wide range of sources and sensors such as social media. Traffic flow sensors cannot know in beforehand if 10 000 people are preparing for a football match. Football matches are random in comparison to events that occur almost daily, such as morning and evening rush hours.

Nevertheless, in this subsection a look is taken at the temporal characteristics of traffic congestions. These are not completely random and can be unveiled by studying traffic flow sensor data. As described in subsection 2.1, ITS provide transportational information: “a traffic congestion of factor 1,5 at 6 P.M. today”. How do they learn to know what traffic congestion means?

It has been a goal to humans to create machines that reason by their own. In literature, this reasoning ability is referred to as “machine learning”, or algorithms that create information from elementary data (Al Najada, Mahgoub, 257, 2016). Here, this information is referred to as causalities, e.g. that time of day affects congestions.

The old view has been that near past influences traffic flow the most. Clearly, a traffic accident (for example) should have a certain “time window”: cleaning the scene shows in traffic flow for several hours onwards. Thus, the first models to predict traffic flow were short-term, involving various moving averages. (Lana et al., 1157, 2016).

Another perhaps more modern view proposes that traffic flow has certain seasonal characteristics. This long-term view reflects a central division in congestion forecasting that follows through this thesis. I.e. there is a typical “Monday of June” or a day that occurs almost same every year. Then, there are “atypical” events (such as accidents) that cause unpredictability. (Lana et al., 1157, 2016).

To find some typical seasonal characteristics (spring/fall -traffic and such) a multi-step algorithm was employed. It was very much like the deep learning algorithm described in 2.2., one fueled by traffic flow sensors’ big data. It clustered traffic flow data sets, and the result was a rough estimate how a “Friday of September” occurs for the most part. (Lana et al., 1159, 2016).

However, another study uses more purposeful terminology to explain these findings. It refers to an "intra-day trend": morning and evening rush hours etc. (Li, Su et al., 294, 2015). In other words, there are cyclical processes that have well-known bases and predictability.

Also, researchers payed attention to the intra-day trend's residual data. These are the deviations from intra-day trend's predictions, and they fall under two categories: Gaussian fluctuations and bursts (Li, Su et al., 295, 2015). I.e. there are patterns that can be roughly quantified with normal distribution (Gaussian fluctuations). Then, there are patterns that can be considered non-predictable. Those reside over 2 σ away from average on the error distribution (bursts).

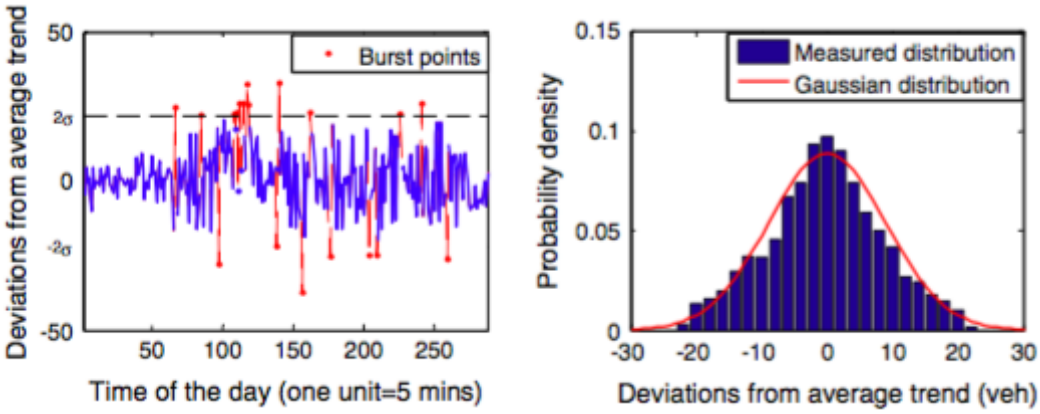

Figure 4. On left: the intra-day trend (in blue) calculated as simple average, with bursts (+/- 2 σ error from the trend) in red. On right: the error distribution after removing the bursts. It follows Gaussian curve roughly. I.e., the intra-day trend is mostly useful for predicting congestions when the rarer events are excluded. Thus, big data sources other than traffic flow sensors (such as social media) must be investigated. (Li, Su, et al., 295, 2015)

Normal distribution can predict recurring day cycles in traffic that behave in Gaussian manner, such as rush hours. However, bursts are spontaneous events that are non-cyclical, such as certain sports matches that occurs rarely. In turn, the algorithm of this thesis must go beyond rush hours to unveil highly dynamic social phenomena. Therefore, in the last chapter of this thesis (4.) a congestion prediction algorithm that uses social media data is presented. It mines big data from sources with few restrictions.

Before that is reviewed what literature answers to the remaining questions. One of those is: have sensors other than traffic flow sensors (ANPR, inductive loop etc.)

ever been utilized in predicting congestions, meaning e.g. social media data, weather etc.?

2.2.2 Social media, an unorthodox source

For starters, it should be noted that social media data does not hold much weight in the literature on congestion analytics by big data. Indeed, only a minority of the roughly 300 articles literature concern it, with the higher emphasis on traffic flow sensors (or other ITS). Thus, from this thesis' standpoint social big data is a slightly unorthodox source.

In the previous subsection, the focus was on the intra-day trend. The data for this trend was collected by traffic flow sensors (Li, Su et al., 298, 2015). In social transportation, a conceptual framework, this data is collected from devices with purpose other than transportation: mobile phones, wearable devices etc. (Chen et al., 2015).

These devices share their data in various social medias. In turn, written language (data) in social media can be mined for traffic related information through Natural Language Processing (NLP) –algorithm (Chen et al., 624, 2015).

NLP processes text through rules that imitate humans' understanding of grammar and vocabulary. Basically, these texts are e.g. the event information that businesses share in social media: sales events, giveaways etc.

Events, for that matter, are assigned a hashtag according to their type, such as: #sales. Consequently, the reviewed big data algorithm could study how (#) sales affect traffic congestions in a specific area, as location can be determined combining the submitted GPS positioning and the written location (Cui, Fu et al., 1551, 2014).

However, the above findings can be criticized. Social data is made by humans, and it is not as objective as the data from traffic flow sensors. It has reservations to both its completeness and quality. Data completeness is defined as how data covers the matter it refers to. An issue with this is demonstrated in the next paragraph.

In social medias, people often post incomplete texts. In a study concerning traffic big data in social medias, NLP was supposed to label information in microblog (Twitter, Weibo etc.) texts. The labels were: incident, time, location etc. Some texts were not informative, although the algorithm was clever to retrieve the missing

information by survey. It could even abandon the microblog if proven unfruitful. (Cui, Fu et al., 1552, 2014)

Then, as an imaginative example of a data quality issue: if one was to mine information in text tagged as (#) congestions, the analysis may lead to error. The big data algorithm may find information by studying #congestions. However, this information is irrelevant if people tag events other than congestions as those. I.e., sentiments affect creation of social media data, and therefore this data is subjective.

While not stellar in completeness or quality, social data captures nuances that other sensors seem not to. For example, a research group found that NLP techniques create information about the feelings of people in transportation (Cui, Fu et al., 1553, 2014). Feelings are subjective information and may require handling by human.

As for traffic congestions, it seems that people find enough issue in those to tweet about (Chen et al., 624, 2015). Thus, social media data provides a prospective source of information to combat traffic congestions. It is not as objective as ITS' traffic flow data, and so it complements the sought composite of big data.

In its subjectivity lies also another asset: drivers define the service rate that means "congestion". Consider e.g. if an often-congested road serves mainly people that are not in a rush. The information is important to the big data algorithm of this thesis, and yet it is not found in traffic flow data. It can be used to market roads on basis of service requirement.

With the service requirements mentioned, the study of prediction algorithm is concluded. So far, this literature reviewed has not presented any real solutions with commercial value. Rather, the purpose was to study the general assumptions about traffic flow. The findings show that congestions are a problem of matching supply to demand. Thus, the commercial solutions preempt congestions by matching service demand to each roads' supply. Those will be studied next.

2.3 Predicting congestions by big data, the modern commercial solutions

Congestions do not seem to be very separate from other urban issues, at least according to the literature. Researchers take a wider standpoint to reducing cars: pollution, bringing transporters to share resources, i.e. a wider social perspective. They propose a big data based incentive system to target commuters according to transport

demand. Allegedly, this system is the first of its kind (Poslad et al., 13070, 2015).

In another model, congestions were described as an integrated issue of a "smart city" -community. In these cities, resource distribution revolves around big data (Mo et al., 2, 2016). I.e., congestions are not a separate commercial interest, but a part of infrastructural solutions for authorities.

The above would explain why most papers on the topic are Chinese. In China, urbanization unlike seen before demands to consider congestions in beforehand to new cities (Mo et al., 1, 2016). Therefore, it would not be wild to assume that the integrated solutions are in high demand as for authorities.

Eventually both research groups point to the same concept: traffic congestions as an integrated issue. They use somewhat similar concepts, and have a common aim to engage citizens in socially and environmentally sustainable urban life. Therefore, they are not both covered here in detail.

However, according to the other traffic congestions are not one-sided: unveiling them requires in-depth analysis of drivers' thoughts. They propose employing big data mining from drivers' IoTs. The analysis's findings reveal that commuters are not very willing to leave from work later than others (Poslad et al., 13090, 2015).

The solution is that incentives for people to commute outside of worst congestions are employed. Test application rewards incentive points should the driver contribute to preempting congestions by e.g. taking foot (Poslad et al., 13085, 2015). Below is a simplified chart of how this is done in the study:

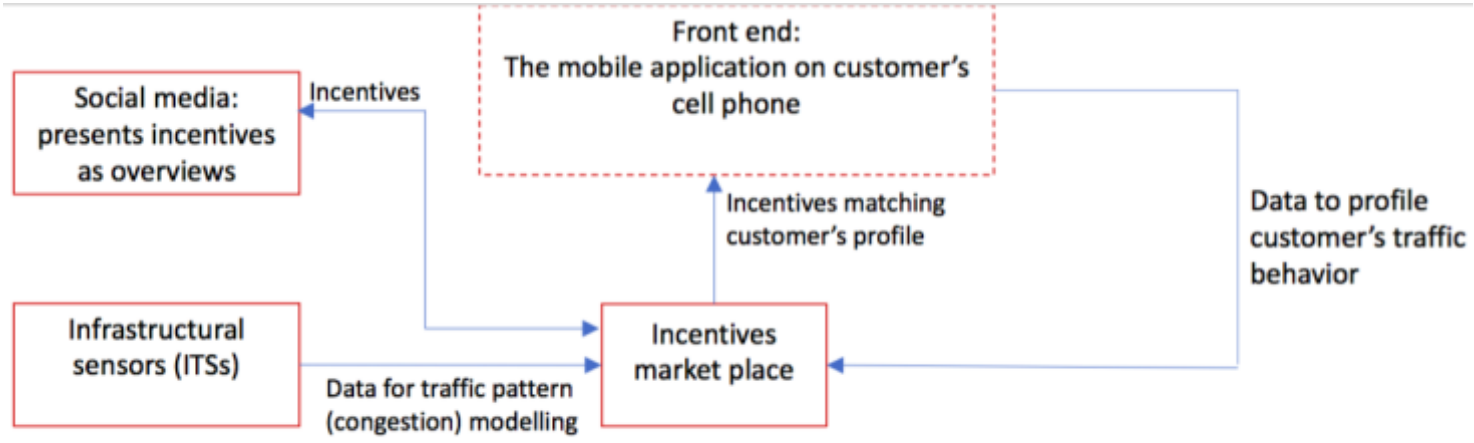

Figure 5. Incentives are distributed to denizens of a "smart city" by drawing data from their IoTs. The data is used to analyze each citizens' behavior in traffic, and then incentives are distributed accordingly. However, the process also requires use of traffic flow sensors (the ITS discussed earlier) for making intra-day trends etc. (Poslad et al., 13081, 2015)

Incentives market place is at the center of the model, drawing big data from various sources. Infrastructural (traffic flow) sensors deliver data for it to make e.g. an intra-day trend, such as the one described in 2.2.1. Customer's mobile phone (IoT) delivers data for customer profiling, and consequently marketing. Thus, congestion forecasting and marketing work together to preempt congestions.

For time being, smart community is one of the few models that could be commercialized. It is a product meant for reducing congestions, and it has been tested. Therefore, it offers some answers whether there is any business worth in predicting traffic congestions by big data, or not. In the sub-section 2.3.1 this business worth and the model's faults are reviewed.

2.3.1 Real-world results of the potential solutions

To reduce congestions, smart city utilizes different data sources and machines. Data from drivers' IoTs is mined for segmentation (rushed, non-rushed etc.), and to map out the "trips" drivers make. Then, incentives are distributed by matching driver's profile to information about daily congestions (provided by traffic flow sensors).

The results of testing smart city model are ambiguous. Commuters' preferences could be mined: 23 % of respondents were interesting in from work later. However, detecting these "trips" proved technically challenging, as many failed to satisfy modality or other required attributes. Thus, only 28 % of the actual trips could be registered as "trips" that affected drivers' earned incentives. (Poslad et al., 13090, 2015)

The test of smart city -model proves that demand for transportation can be mined. It also proves that some people are willing to prevent congestions by changing their behavior. If the future test runs of the model prove this again, then congestions can probably be prevented to some extent. However, it is too early to talk about any monetary worth that the model could have.

The concept of building a big data fueled (smart) community is unarguably fanciful. It deserves applauds for its approach proposing that traffic congestions are structural: tied to roads upon building them. However, in this fancifulness lies also places for criticism.

For starters, smart cities are interconnected and dependent on their costly machines. As a contrast, big data is a technology (common knowledge) and hence independent of infrastructure: computations can be done anywhere in the world. In turn, smart city is a machine dependent community built for the sole purpose of applying big data. Therefore, it is highly vulnerable to power breakdowns, natural disasters etc. (Mo et al., 6, 2016).

What this means for congestion prediction is that the big data solution should be relatively light: universally applicable with low requirements for infrastructure. This is at least if one is to build commercial solution to catering all drivers around the world.

It has been reviewed what literature says on predicting traffic congestions by big data. The field is young with barely 300 articles of which many are yet to be accepted in scientific journals. There is much debate on the viability of different methods and their coverage of big data. Thus, the literature concludes that the field is commercially interesting, as the potentials are likely yet mostly unexplored.

3. Conclusions

3.1 Studies' methods, imperfections & justifications

The literature presents multiple perspectives to predicting traffic congestions with big data. However, each study is often limited to just few data sources. However, there is a central division between sensors: others are maneuverable, while others are embedded. I.e., infrastructure embedded traffic flow sensors tell about traffic on roads that they are sited on, whereas drivers' cell phones (IoT) can be mined for data anywhere.

IoTs benefits to traffic flow sensors are considerable in many instances. After all, traffic flow sensors may be costly to install in number. For comparison, GPS based techs such as Google Traffic require no such infrastructure to be installed. Yet, even those do not seem to provide sufficient coverage as discussed in 1.3.

With the 10-years old solution being addressed, some clarifications are needed. First off, GPS does not equal a car. I.e., GPS may not provide objective road traffic data: not everyone having it is a driver, and not every driver transmits positioning, although nearly so in Finland as of 2016's developments.

To address this, the Harbin research group had a special source: GPS mined straight from the navigators embedded in taxis. While the source is objective it is also discriminative: geographical limitations such as the taxi prohibited areas (discussed in 1.3.). In some areas, limitations such as these may hamper the view of the local congestions.

Then, as found in a study seeking sites for e-car recharge stations, analysis results of taxi GPS data cannot be generalized to private cars (Cai et al., 45, 2014). Reasons may vary from income level to taxis being the minority etc. However, the point remains: big data is not about finding an ultimate data source. It is about seeking a composite that sees past the biases in separate data sources.

Therefore, instead of bashing GPS-, or traffic flow sensor-, or social media data, in this thesis they have been all considered. The conclusions section should reflect this. It seems that there should be intra-day trend for the common days. Then, for the other part of predictive analytics should be employed e.g. social media data.

Some researchers were interested in the congestions' intra-days. They found the best source material to this as ITS' traffic flow data. Traffic flow data is objective, yet naive analysis could lead to underestimating rare events. However, researchers were critical to describe this trade-off: complexity versus performance (Li, Su et al., 305, 2015).

What this trade-off probably means is that the intraday trend provided satisfying yet imperfect prediction. Complexity would be to add variables, and if every variable behind congestions could be added then there would not be "bursts" (severe deviations). However, the idea was to predict congestions efficiently rather than precisely.

Therefore, the method was justified in computational terms: adding all data to the model would require huge capacity and would cost tons (Li, Su et al., 292, 2015). In turn, this means that sources of big data should not be added on a naive "more is more" –basis. Instead, it is all contextual.

On the other hand, ITS data may not be the best to figure out contextual matters in traffic. In all the text data in social medias are information that ITS fail to understand. Researchers employed an NLP algorithm to find these contextual "feelings" (Cui, Fu et al., 1553, 2015). This perspective complements the quantitative one given by the ITS.

To demonstrate: congestion costs may in some instances be distributed

unevenly between road users. Consider for instance 200 people about to miss a flight stuck in a jam of 77 000 commuters. Information given by NLP is to identify roads that are most sensitive to the commercial losses of congestions.

Thus, the service rate of “congested road” may not be the same on every road. Roads filled often by masses of people not in a rush must have lower tolerance for congestions. Otherwise people in rush get directed to these lanes, and cause even worse congestions. Therefore, whether congestions incur losses on certain roads depends on those roads’ users, e.g.: “#missedflight due to #congestion, #unhappy”.

To add, if congestions are not normal distributed, then probabilistically the lion-share of costs may be caused by the more severe ones. This is to argue that what was called “efficient” in the intra-day trend is not, although on average days it performs very well with low computations.

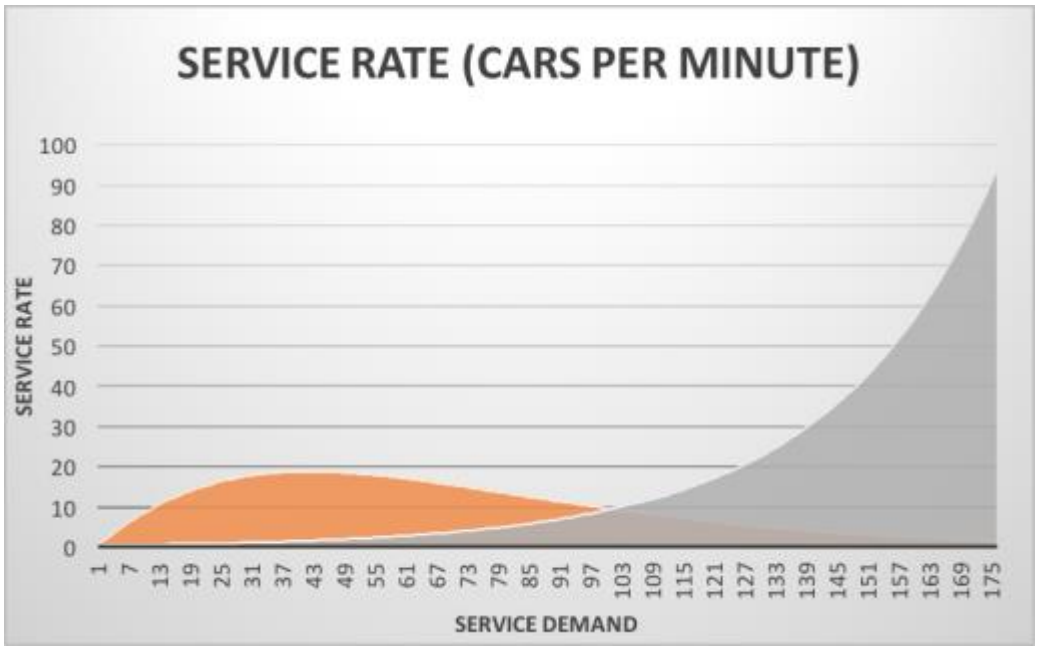

Figure 6. The above graph complements the one in 1.3 by showing that lost service (and in turn costs) (grey area) rise in convex trend the more severe and rarer the congestion. If the probability density of these is non-Gaussian then they may dominate the average days in costs.

To conclude this thesis, in the next sub-section a route guidance system in driver’s car is demonstrated. Its real-time congestion predictions are fueled with big data. This big data algorithm understands not only intra-day trends, but also contextual matters. As social media posts are usually filled with information on places, time, type of occurrence etc., it mines the data of those to fill out gaps in the intraday-trend.

3.2 Big data enhanced traffic congestion prediction algorithm

The big data solution to predict congestions should probably have regional teams that develop the algorithm. This is due to contextual differences between different major traffic areas. Traffic prediction models are not efficient if they have generic lists of variables in their models. These may lead to unnecessary computations or models that are heavy to run on computer, as discussed in the previous subsection (3.1).

Therefore, separate teams handle the development of the area specific algorithm, of which predictions are rendered real-time through cloud on driver's screen, a solution known as dynamic route guidance –system. As the base of these predictions are the stable Gaussian intra-day trends. As explained in 2.2.1, these trends were made up of data mined from ITS as a simple average.

The algorithm has also another source of big data: social media. This is for the reasons given in 2.2.2. First, there are contextual matters that need to be considered to sharpen the definition of traffic congestions. Then, social phenomena cannot be captured by traffic flow sensors. The social media part of the algorithm is discussed in the next subsection (3.2.1).

3.2.1 Social big data's role in congestion prediction

ITS data fails to capture all the dynamicity in traffic. After all, the prediction may be a simple average of longer time span such as 1000 cars on Mondays at between 9-10 PM. However, there are matters that occur over time that disregard traffic. Consider e.g. a hugely popular game show that draws such many people on front of TV that it affects Friday nights' traffic flows.

Quite possibly it is those events that the study on intra-day trend referred to as "exogenous bursts" that are "time-variant" (Li, Su et al., 293, 2015). This implies the reality that traffic congestions are much unpredictable for spontaneous events. Thus, the algorithm to predict them is left with the imagination of its developers.

While probably unpredictable, a congestion prediction algorithm is demonstrated here that attempts to fill-in the intra-day's gaps. Congestion-explaining factors in certain areas may be e.g. time-variant events such as decade-specific TV-

shows. Some popular ones in Finland such as Putous air only during fall and winter, and the intra-day trend may smooth these out.

Then, completely random events such as “classic football matches” or “Bruce Springsteen concerts” in metropolitan suburbs are smoothed out with the most severe consequences. From the standpoint of this literature review, the respective traffic congestion may pass the 2 σ error limit proposed for “bursts” (Li, Su et al., 293, 2015).

Consider for instance, a Champions’ League football team. There is high variability to the outcomes of the season and to matches’ popularity. Consequently, there is also high variability to how much people post about the matches in social media (#footballmatch). Adding all these variables to a prediction model, e.g. a certain kind of match, would quickly add to its complexity.

Thus, something is needed that tags events without drawing data from many different sources such as promoters’ ticket sales. After all, ticket sales reflect just one aspect of an event. But, both those who go on stadium and watch the event from sports bar post in social media with hashtags: #footballmatch #ChampionsLeagueFinal2017 #Cardiff.

In social media, every tag is equal and so it offers a perspective to compare each (#) football matches impact on congestions. Nevertheless, first a causality must be established as was in the case of the intra-day trend. To do so, the researchers employed a Granger causality test (Li, Su et al., 295, 2015).

The test is a simple regression to study links between matters. Technically speaking, it tests if football matches (X) and congestions (Y) predict the latter better than congestions (Y) alone do. It is not claimed that it is the right one to study this precise link. The reason for bringing Granger causality up here is for pure demonstration.

Another reservation is that there might not be enough empiric material to study the link between e.g. congestions and football matches. As the sample pool could turn out small, so could the error distribution of predictions turn out volatile. As pointed out, intra-day trend generates high relative error on roads that service just few people (Lv et al., 2015).

Nevertheless, in the next subsection a simple algorithm is demonstrated that focuses on filling in the intra-day model’s gaps (bursts): a process description and its flowchart.

3.2.2 An example algorithm to predict traffic congestions by big data

1. Algorithm mines hashtags that indicate (eg.) upcoming football matches (#footballmatch). It mines the time of event through NLP as “19:30” and assigns it to this exact event. It also mines its location: Strobelallee 50 44139 Dortmund, Germany. Location links the event to the correct traffic area.
2. Algorithm performs a causal test to #footballmatch to decide whether there is a causality between them and congestions. This is done by running Granger regression between the time series of #footballmatch and the time series of traffic flow in Dortmund.
3. The results show that football matches Granger cause traffic congestions in the Dortmund area.
4. If the causal test is passed (with certain requirements), then algorithm decides that (#) football matches are a variable in its congestion prediction model.
5. In the future, each time a football matches occurs, driver’s route guidance - system delivers a more precise congestion forecast. This is done with the regression model based on past data.

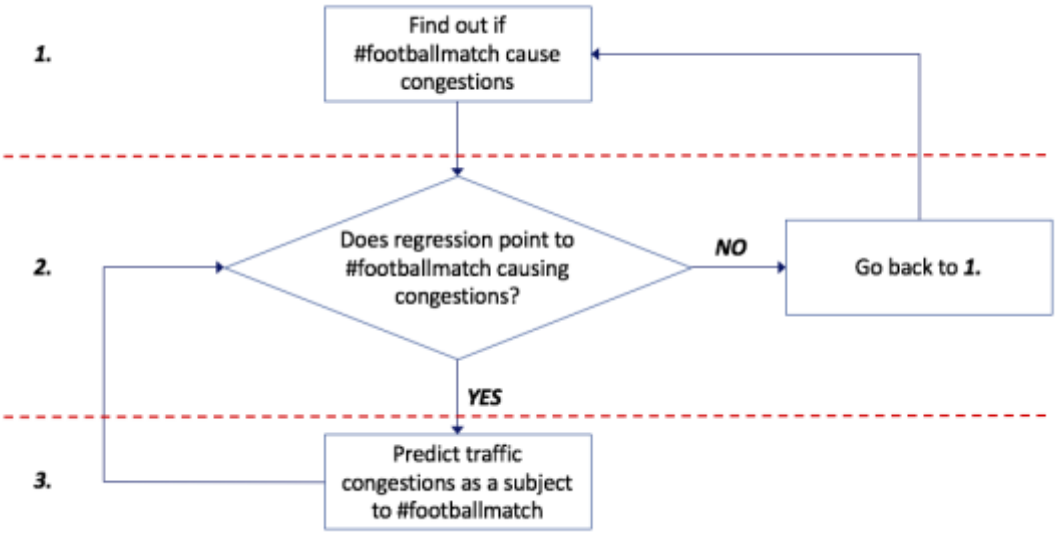

Figure 7. The flowchart of the described process.

3.2.3 Conclusive words on traffic congestions and big data

Each source of traffic related big data for congestion predictions offers a different insight. Indeed, there are not many solutions presented that integrate many sources. However, those who proposed smart community model recognized the need for synchronized use of infrastructural sensors (much like ITS) and social media.

Thus, instead of saying that social media (or some other) is a necessary source, it is concluded here that "congestion predictions by big data" is a justified expression when:

1. It is recognized that separate big data sources may be biased. Therefore, big data may require utilizing more than one data source. This selection is based on contextual matters.
2. Statistical methods are used to consider each prospective source. These may vary from e.g. Granger causality to other regression models.
3. Big data fueled predictions are efficient, meaning that they can be machine run in real time without human workforce. Thus, they may need to be enhanced by techs such as IoT, cloud computing etc.
4. Predictive solutions are light in infrastructure, meaning that they are universally applicable without many machines embedded in the infrastructure. =>
5. Traffic congestion forecasts by big data can be effectively made into a commercial product.

REFERENCES

Abdel-Aty, M., Shi, Q. 2015. *Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways, Transportation Research Part C,* vol 58, pages 380-394

Al Najada, H., Mahgoub, I. 2016. *Big vehicular traffic data mining: Towards accident and congestion prevention*, *International Wireless Communications and Mobile Computing Conference (12th*), pages 256-261

Cai, H., Chiu, A., Hu, X., Jia, X., Xu, M. 2014. *Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet, Transportation Research Part D,* vol 33, pages 39-46

Capra, L., Wolfson, O., Yang, H., Zheng, Y. 2014. *Urban computing: Concepts, methodologies, and applications, ACM Transactions on Intelligent Systems and Technology*, vol. 5, article 38, pages 1-55

Chen, S., Chen, W., Shen, D., Wang, P., Wang, X., Yang, L., Zhang, Q., Zheng, X. 2016. *Big Data for Social Transportation*, *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, pages 620-630

Cools, M., Cui, J., Hu, K., Janssens, D., Liu, F., Wets, G. 2016. *Identifying mismatch between urban travel demand and transport network services using GPS data: A case study in the fast-growing Chinese city of Harbin*, vol. 181, pages 4-18

Cui, J., Dong, C., Fu, R., Zhang, Z. 2014. *Extraction of traffic information from social media interactions: Methods and experiments, IEEE International Conference on Intelligent Transportation Systems (17th),* pages 1549-1554

Del Ser, J., Lana, I., Olabarrieta, I. 2016. *Understanding daily mobility patterns in urban road networks using traffic flow analytics, Proceedings of the NOMS 2016,* pages 1157-1162

Dogdu, E., Kucukayan, G., Ozbayoglu, M. 2016. *A real-time autonomous highway accident detection model based on big data processing and computational intelligence, IEEE International Conference on Big Data (4th)*, pages 1807-1813

Duan, Y., Kang, W., Li, Z., Lv, Y., Wang, F.-Y. 2015. *Traffic Flow Prediction with Big Data: A Deep Learning Approach*, *IEEE Transactions on Intelligent Transportation Systems,* vol. 16, pages 865-873

INRIX. 2014. *Annual cost of gridlock in Europe and the US will increase 50 percent on average to $293 billion by 2030*, press release (14 October)

Li, T., 2005. *Nonlinear dynamics of traffic jams, Physica D: Nonlinear Phenomena,* vol. 207, pages 41-51

Li, L., Li, Y., Li, Z., Lin, Y., Su, X., Wang, Y. 2015. *Robust causal dependence mining in big data network and its applications to traffic flow predictions, Transportation Research Part C: Emerging Technologies,* vol. 58, pages 292-307

Ma, A., Mei, H., Poslad, S., Wang, Z. 2015. *Using a smart city IOT to incentivize and target shifts in mobility behavior - Is it a piece of pie?, Sensors (Switzerland),* vol. 15, pages 13069-13096

Mo, Z., Peng, Y., Shen, J., Wu, Y., Zhang, W. 2016. *Smart city with Chinese characteristics against the background of big data: Idea, action and risk, Journal of Cleaner Production*, pages 1-7